MicroscopyCLIP: A Domain-Specific Vision-Language Model for Optical Microscopy

Zhuoqin Yang[1,2], Jiansong Zhang[1], Xiaojun Wang[3], Xiaoling Luo[1], Xiaofei Huang[1], Jie Wang[2], Kian Ming Lim[2], Zheng Lu[2(✉)], and Linlin Shen[4,5(✉)]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2] School of Computer Science, University of Nottingham Ningbo China, Ningbo, China
[3] Department of Obstetrics and Gynecology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou, China
[4] School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[5] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, China
✉ Corresponding authors: zheng.lu@nottingham.edu.cn, llshen@szu.edu.cn

Abstract. Microscopy images provide fine-grained visual evidence of cellular morphology and are fundamental to biomedical research and clinical diagnosis. However, existing vision-language models, including CLIP and its medical variants, are primarily trained on natural images, radiology data, or medical literature, and therefore lack effective semantic alignment for microscopy imagery, limiting their performance in microscopy zero-shot recognition and cross-dataset transfer capability. In this paper, we propose MicroscopyCLIP, a domain-specific vision-language model tailored for microscopy image understanding. We construct a large-scale microscopy image-text dataset consisting of approximately 210K fluorescence microscopy images paired with textual descriptions, followed by systematic expert verification to ensure semantic consistency. Based on this dataset, we perform full-parameter fine-tuning of a pretrained CLIP model to realign the vision-language embedding space for microscopy data. Experiments demonstrate that MicroscopyCLIP substantially improves the zero-shot classification performance for microscopy images. On a held-out 8-class kidney cell benchmark, it achieves 66.17% Top-1 accuracy under prompt-based zero-shot inference without task-specific training on the target dataset. Across six external microscopy benchmarks, MicroscopyCLIP consistently outperforms existing vision-language models in both zero-shot and linear probing settings, indicating robust cross-dataset generalization. Code: https://github.com/SeriYann/MicroscopyCLIP

Keywords: Microscopy Image Analysis · Contrastive Multimodal Learning · Vision-Language Pretraining.

1 Introduction

Microscopy images are fundamental to biomedical research and clinical diagnosis, as they provide fine-grained visual evidence of cellular morphology and pathological states [9]. Through stained tissue sections observed under optical microscopes, pathologists assess critical morphological features such as cell size, spatial organization, cytoplasmic characteristics, and nuclear atypia, which inform tumor grading, inflammatory assessment, and other histopathological evaluations [12,18]. Beyond traditional workflows, advances in medical artificial intelligence have supported image-based diagnosis, while microscopy images have become a core data source for digital pathology and automated cell classification [1,22,25].

Meanwhile, contrastive vision-language models, particularly CLIP [16], have demonstrated remarkable zero-shot recognition and transfer capabilities in natural image domains by aligning images and texts within a shared semantic space [6,26,27]. Inspired by this success, several medical CLIP variants, including MedCLIP [23] and PMC-CLIP [8], have extended contrastive learning to medical imaging tasks using radiology images, clinical reports, and medical literature [7,13]. However, microscopy imagery remains relatively underrepresented in existing medical vision-language models. Current approaches are not tailored to the fine-grained morphological characteristics inherent in microscopic images, limiting their applicability to microscopy-based zero-shot recognition and cross-dataset generalization.

A key reason for this gap lies in the nature of microscopy image annotations: unlike radiological images, which are typically accompanied by structured diagnostic reports, microscopy images are more reliant on physicians' subjective observations and are seldom documented using standardized textual formats [17]. Even in some public datasets where microscopy images are labeled, such annotations are usually coarse-grained and weakly supervised, offering limited utility for training image-text alignment models [14]. This significantly hinders the direct generalization of existing CLIP models to microscopy tasks and restricts their potential in zero-shot classification and multimodal retrieval.

To address these challenges, we propose MicroscopyCLIP, a domain-adaptive contrastive vision-language model specifically designed for microscopy image understanding. To facilitate effective image-text alignment in the microscopy domain, we construct a large-scale microscopy image-text dataset consisting of 207,988 optical microscopy images of human kidney cells paired with expert-reviewed and refined textual descriptions. Based on this dataset, we perform full-parameter fine-tuning of a pretrained CLIP model to realign the shared embedding space toward fine-grained cellular morphology.

We comprehensively evaluate MicroscopyCLIP under both in-domain and cross-dataset settings, including zero-shot classification and label-efficient adaptation. On a held-out kidney microscopy test split, MicroscopyCLIP achieves 66.17% Top-1 zero-shot accuracy without using any downstream labels, substantially outperforming existing vision-language baselines. We further assess cross-dataset transfer on six public microscopy benchmarks, where MicroscopyCLIP consistently improves zero-shot performance via prompt-based inference,

and also delivers strong results under linear probing by freezing the image encoder and training only a linear classifier, demonstrating robust generalization in optical microscopy. Our main contributions are as follows:

- We identify a critical gap in existing medical vision-language models, namely the lack of domain-specific alignment for fine-grained microscopy imagery, and introduce MicroscopyCLIP to address this limitation through domain-adaptive contrastive learning.
- To enable effective image-text alignment in the microscopy domain, we construct a large-scale microscopy image-text dataset comprising 207,988 fluorescence microscopy images paired with expert-verified textual descriptions, providing structured supervision for fine-grained cellular semantics.
- We conduct comprehensive evaluations under both in-domain and cross-dataset settings, demonstrating that MicroscopyCLIP substantially improves zero-shot recognition and transfer performance across multiple microscopy benchmarks, validating the effectiveness of contrastive vision-language learning in microscopy domain.

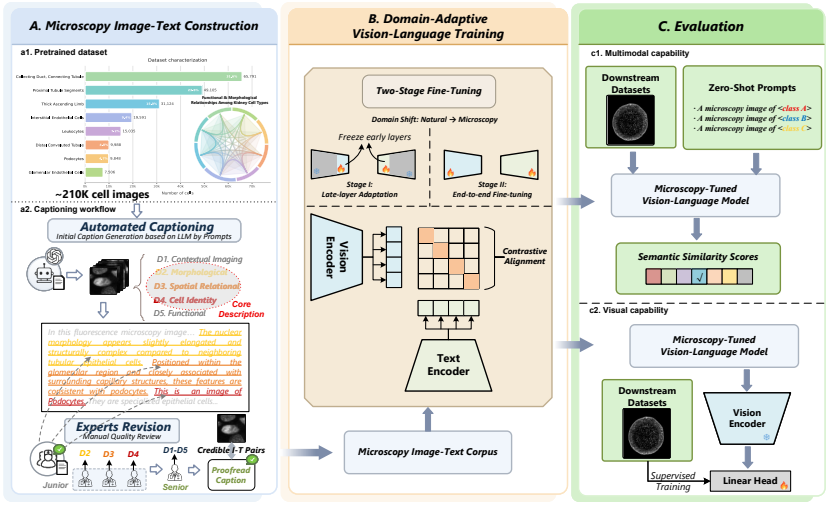

Fig. 1. Overview of the proposed MicroscopyCLIP framework. A: Microscopy image-text supervision is constructed via LLM-based caption generation followed by expert verification. B: A pretrained CLIP model is domain-adaptively fine-tuned using a two-stage progressive strategy to align multimodal representations with microscopy-specific semantics. C: The resulting microscopy-tuned model is evaluated under zero-shot classification using prompt-based inference and linear probing on downstream datasets.

2 Method

2.1 Microscopy Image-Text Supervision via LLM Captioning and Expert Verification

Microscopy images are rarely accompanied by standardized textual descriptions, limiting the direct application of vision-language pretraining methods. To address this limitation, we construct microscopy-specific image-text supervision through a structured pipeline combining LLM-based caption generation with hierarchical medical expert verification.

We curate 207,988 grayscale fluorescence microscopy images spanning eight kidney cell types from publicly available datasets [10,24]. This kidney cortex collection is one of the largest publicly available cell-level optical microscopy datasets, providing clear cellular identity labels and diverse morphological patterns. Initial captions are generated using structured prompts emphasizing contextual (D1), morphological (D2), spatial (D3), identity (D4), and functional (D5) components.

To ensure semantic accuracy, we establish a hierarchical human review protocol involving medical researchers with doctoral-level expertise. Junior reviewers primarily refine morphology (D2), spatial relationships (D3), and identity expression (D4), where visual-text alignment is most critical. Manual revisions focus on refining nuclear morphology descriptions, clarifying spatial localization, removing unsupported visual or spatial inferences, and normalizing cell-type terminology. A senior expert subsequently performs an integrative review across all semantic components (D1-D5) to ensure overall medical consistency before forming the final image-text pairs. Through this manual verification process, we obtain a high-quality microscopy image-text corpus that provides morphology-aligned supervision for contrastive vision-language training.

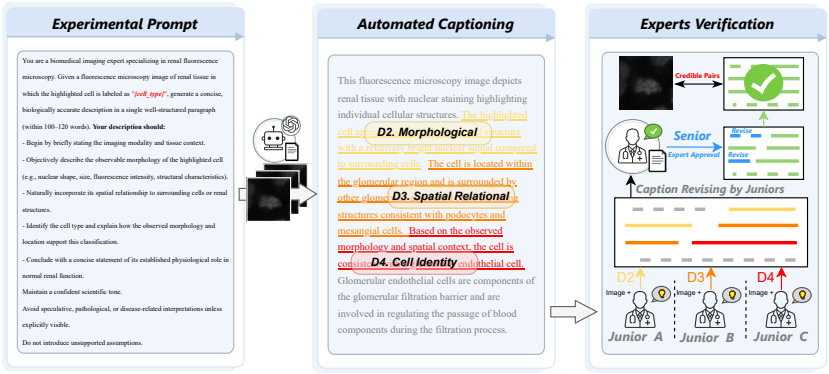

Fig. 2. Structured pipeline for microscopy image-text supervision construction. LLM-generated captions are refined through multi-level expert review to ensure medical accuracy and visual consistency.

2.2 Domain-Adaptive Vision-Language Training

Using the verified microscopy image-text pairs described above, we perform domain-adaptive vision-language fine-tuning to align pretrained multimodal representations with microscopy-specific semantics. MicroscopyCLIP is initialized from the CLIP ViT-B/32 model, consisting of an image encoder $f_\theta(\cdot)$ and a text encoder $g_\phi(\cdot)$. Given an image-caption pair (I, T) from the verified corpus, normalized embeddings are computed as:

$$\mathbf{v} = \frac{f_\theta(I)}{|f_\theta(I)|}, \quad \mathbf{t} = \frac{g_\phi(T)}{|g_\phi(T)|}. \tag{1}$$

For each minibatch of paired microscopy images and captions, we retain the standard CLIP contrastive objective by optimizing a symmetric image-to-text and text-to-image loss:

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{I\to T} + \mathcal{L}_{T\to I}\right), \tag{2}$$

where both terms are implemented as cross-entropy losses over temperature-scaled image-text similarity logits. This preserves the original multimodal alignment objective while enabling domain-specific adaptation through microscopy supervision.

2.3 Two-Stage Progressive Full-Parameter Fine-Tuning

Although the standard contrastive objective enables domain adaptation, direct end-to-end fine-tuning under substantial domain shift between natural image pretraining data and microscopy morphology may lead to unstable optimization [4]. Early layers may potentially drift rapidly, leading to suboptimal feature reuse and slow domain adaptation. To stabilize representation adaptation while enabling full cross-modal realignment, we adopt a two-stage progressive training strategy.

(1) **Stage I (late-layer adaptation).** We first freeze early Transformer blocks in both vision and text encoders and update only late blocks together with projection layers. This stage primarily adapts high-level semantic representations using a relatively larger learning rate while preserving transferable low-level visual features learned during large-scale pretraining.

(2) **Stage II (end-to-end fine-tuning).** We then unfreeze all layers and perform full-parameter fine-tuning to fully align microscopy visual representations with verified textual semantics using a smaller learning rate. In practice, Stage I and Stage II are trained using different learning rate scales. Training is performed for up to 30 epochs using the AdamW optimizer [11].

Fig. 3. Performance comparison across six downstream datasets. (a) Zero-shot evaluation (dashed: vanilla, solid: with CORAL). (b) Linear probing with frozen image encoder.

3 Evaluation

3.1 Experimental Setup

We compare MicroscopyCLIP with representative vision-language models, including CLIP [16], MedCLIP [23], BioMedCLIP [28], and PMC-CLIP [8], covering both general-domain and biomedical-domain multimodal pretraining paradigms.

The evaluation is primarily conducted under two classification protocols. First, we assess prompt-based zero-shot classification using prompt-guided text embeddings without training task-specific classifiers. Second, we evaluate representation transfer via linear probing, where a linear classifier is trained on frozen image encoder features. Since the downstream microscopy benchmarks do not provide paired image-text annotations, classification-based evaluation serves as the primary protocol for assessing cross-dataset generalization.

In addition, we conduct cross-modal retrieval evaluation on the held-out kidney dataset using paired image-text annotations to further assess image-text alignment quality. Performance is reported using Top-1 classification accuracy for classification tasks and Recall@10 (R@10) for retrieval tasks. All compared models are evaluated under the same prompt templates and training protocols for fair comparison. All experiments were conducted on four NVIDIA RTX A6000 GPUs.

3.2 In-Domain Zero-Shot Evaluation

We evaluate zero-shot recognition on held-out kidney microscopy images from the same domain using the prompt template *"A microscopy image of {class}."* MicroscopyCLIP achieves 66.17% Top-1 accuracy on the 8-class kidney cell

Table 1. Cross-dataset zero-shot and linear probe performance (Top-1 accuracy, %). Red values indicate absolute improvements over the best baseline.

Dataset	CLIP	MedCLIP	BioMedCLIP	PMC-CLIP	**MicroscopyCLIP (Ours)**
Zero-shot Classification					
Kidney (Held-out)	10.68	14.94	7.04	9.78	**66.17**(+51.23)
CAD Cells-M	50.00	50.00	43.10	43.10	**50.00**(—)
CAD Cells-A	9.87	4.61	7.45	16.67	**21.71**(+5.04)
Jurkat Cell Cycle	4.63	2.09	2.09	2.09	**32.26**(+27.99)
C. elegans Infection	56.67	46.67	42.22	53.33	**60.00**(+3.33)
C. elegans Metabolism	40.00	42.00	40.00	42.20	**43.55**(+1.55)
MCF7 / A549 Cytoplasm	75.00	75.00	75.00	65.63	**79.68**(+4.68)
Linear Probe Classification					
CAD Cells-M	82.35	82.35	76.47	88.24	**94.12**(+5.88)
CAD Cells-A	62.50	56.25	50.00	68.75	**81.25**(+12.50)
Jurkat Cell Cycle	83.33	85.19	83.33	87.96	**88.89**(+0.93)
C. elegans Infection	70.83	72.22	68.05	73.61	**76.39**(+2.78)
C. elegans Metabolism	50.00	50.00	40.00	60.00	**60.00**(—)
MCF7 / A549 Cytoplasm	94.74	96.55	94.74	97.37	**98.28**(+0.91)

benchmark, compared with 10.68% for CLIP, 14.94% for MedCLIP, 7.04% for BioMedCLIP, and 9.78% for PMC-CLIP. These results demonstrate that domain-adaptive vision-language alignment significantly improves zero-shot recognition performance in microscopy. The large performance gap suggests that existing biomedical vision-language models, primarily trained on radiology images and clinical reports, may have limited exposure to microscopy morphology, highlighting the importance of modality-specific supervision.

To further assess image-text alignment quality, we additionally conduct cross-modal retrieval on the held-out kidney dataset. As shown in Table 2, MicroscopyCLIP consistently outperforms CLIP and MedCLIP in both image-to-text and text-to-image retrieval, indicating improved cross-modal semantic alignment for microscopy imagery.

Table 2. Cross-modal retrieval performance on the held-out kidney dataset (R@10, %).

Model	Image→Text	Text→Image
CLIP	6.1	5.5
MedCLIP	4.5	5.1
MicroscopyCLIP	**22.5**	**18.9**

3.3 Cross-Dataset Generalization

To assess generalization beyond the kidney pretraining distribution, we evaluate MicroscopyCLIP on six external microscopy classification datasets spanning diverse cell types, organisms, and experimental conditions, including Murine Cath.a differentiated (CAD) cells [2], Cell Cycle Jurkat Cells [3], C. elegans

infection marker and metabolism assay [15,21], and Human MCF7 and A549 cytoplasm [10].

Table 1 reports the primary quantitative comparison using raw zero-shot and linear probing results, reflecting direct cross-dataset transfer performance without target-specific adaptation. Across all external datasets, MicroscopyCLIP consistently achieves the strongest zero-shot performance, indicating improved cross-dataset semantic alignment of the learned vision-language embeddings.

Zero-shot inference relies purely on embedding similarity and is therefore more sensitive to distribution shifts [20]. To further analyze the effect of cross-dataset covariance mismatch, Figure 3 (a) additionally presents zero-shot performance with feature-level Correlation Alignment (CORAL) [19] applied at inference time. CORAL performs second-order statistical alignment and is applied uniformly across all compared models as a post-hoc analysis rather than a training component. While all models exhibit moderate improvements after alignment, MicroscopyCLIP maintains the strongest performance and demonstrates larger gains on several datasets. This suggests that the learned representations are not only semantically aligned but also structurally more amenable to cross-dataset statistical alignment. These results demonstrate that domain-adaptive vision-language alignment improves robustness to cross-dataset variability in microscopy imaging.

3.4 Ablation Study

We conduct ablation experiments to evaluate the impact of the training strategy and prompt design. We compare the proposed two-stage progressive fine-tuning strategy with single-stage full fine-tuning and parameter-efficient LoRA-based adaptation [5]. As shown in Table 3, the two-stage strategy achieves the best performance on both the in-domain kidney benchmark and the external Jurkat Cell Cycle dataset. Single-stage full fine-tuning yields slightly lower accuracy, while LoRA results in a substantial drop in performance under the same training budget. This suggests that parameter-efficient adaptation may be insufficient to fully realign representations under strong microscopy domain shift, whereas progressive full-parameter adaptation better supports stable domain adaptation.

We further assess robustness to prompt formulation by testing multiple text templates, including generic image descriptions and microscopy-specific prompts. Performance remains relatively stable across different formulations, indicating that the learned embedding space captures domain-level semantic alignment rather than relying heavily on specific lexical patterns.

4 Conclusion

We present MicroscopyCLIP, a domain-adaptive vision-language framework tailored for microscopy image understanding. By constructing expert-verified microscopy image-text supervision and adopting a progressive full-parameter adaptation strategy, our approach effectively aligns pretrained multimodal representations with microscopy-specific semantics. Extensive experiments demonstrate

Table 3. Ablation study on in-domain held-out Kidney dataset and downstream dataset Jurkat Cell Cycle zero-shot benchmarks (Top-1 Accuracy, %).

Setting	Kidney	Jurkat
Training Strategy		
Single-stage Full Fine-tuning	62.33	28.91
LoRA Fine-tuning	27.69	5.31
Two-stage Progressive Fine-tuning	**66.17**	**32.26**
Prompt Template		
This is a {class}.	65.05	30.03
An image of {class}.	65.44	31.28
A microscopy image of {class}.	66.17	32.26
An optical microscopy image showing {class}.	65.73	32.68

substantial improvements in zero-shot recognition and consistent cross-dataset generalization across diverse microscopy benchmarks. Future work will explore stronger domain alignment mechanisms and broader multimodal supervision to improve cross-domain robustness and extend the framework to additional microscopy modalities.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62576216 and Guangdong Provincial Key Laboratory under Grant 2023B1212060076.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J.: Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature medicine **25**(8), 1301–1309 (2019)
2. Edwards, P., Skruber, K., Milićević, N., Heidings, J.B., Read, T.A., Bubenik, P., Vitriol, E.A.: Tdaexplore: Quantitative analysis of fluorescence microscopy images through topology-based machine learning. Patterns **2**(11) (2021)
3. Eulenberg, P., Köhler, N., Blasi, T., Filby, A., Carpenter, A.E., Rees, P., Theis, F.J., Wolf, F.A.: Reconstructing cell cycle and disease progression using deep learning. Nature communications **8**(1), 463 (2017)
4. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
5. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
6. Huang, X., Gong, H.: A dual-attention learning network with word and sentence embedding for medical visual question answering. IEEE Transactions on Medical Imaging **43**(2), 832–845 (2023)
7. Jang, J., Kyung, D., Kim, S.H., Lee, H., Bae, K., Choi, E.: Significantly improving zero-shot x-ray pathology classification via fine-tuning pre-trained image-text encoders. Scientific Reports **14**(1), 23199 (2024)

8. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
9. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)
10. Ljosa, V., Sokolnicki, K.L., Carpenter, A.E.: Annotated high-throughput microscopy image sets for validation. Nature methods **9**(7), 637 (2012)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Lu, A.X., Kraus, O.Z., Cooper, S., Moses, A.M.: Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting. PLoS computational biology **15**(9), e1007348 (2019)
13. Lu, Z., Li, H., Parikh, N.A., Dillman, J.R., He, L.: Radclip: Enhancing radiologic image analysis through contrastive language–image pretraining. IEEE Transactions on Neural Networks and Learning Systems (2025)
14. Ma, J., Xie, R., Ayyadhury, S., Ge, C., Gupta, A., Gupta, R., Gu, S., Zhang, Y., Lee, G., Kim, J., et al.: The multimodality cell segmentation challenge: toward universal solutions. Nature methods **21**(6), 1103–1113 (2024)
15. O'Rourke, E.J., Soukas, A.A., Carr, C.E., Ruvkun, G.: C. elegans major fats are stored in vesicles distinct from lysosome-related organelles. Cell metabolism **10**(5), 430–435 (2009)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
17. Rädsch, T., Reinke, A., Weru, V., Tizabi, M.D., Schreck, N., Kavur, A.E., Pekdemir, B., Roß, T., Kopp-Schneider, A., Maier-Hein, L.: Labelling instructions matter in biomedical image analysis. Nature Machine Intelligence **5**(3), 273–283 (2023)
18. Schmidt, U., Weigert, M., Broaddus, C., Myers, G.: Cell detection with star-convex polygons. In: International conference on medical image computing and computer-assisted intervention. pp. 265–273. Springer (2018)
19. Sun, B., Feng, J., Saenko, K.: Return of frustratingly easy domain adaptation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 30 (2016)
20. Tu, W., Deng, W., Gedeon, T.: A closer look at the robustness of contrastive language-image pre-training (clip). Advances in Neural Information Processing Systems **36**, 13678–13691 (2023)
21. Wählby, C., Kamentsky, L., Liu, Z.H., Riklin-Raviv, T., Conery, A.L., O'rourke, E.J., Sokolnicki, K.L., Visvikis, O., Ljosa, V., Irazoqui, J.E., et al.: An image analysis toolbox for high-throughput c. elegans assays. Nature methods **9**(7), 714–716 (2012)
22. Wang, J., Liu, M., Liao, H., Fan, J., Zhu, H., Ma, T., Yang, D., Ni, F., Zhang, F., Jin, G., et al.: Development of deep learning-based narrow-band imaging endocytoscopic classification for predicting colorectal lesions from a retrospective study. Nature Communications **16**(1), 8351 (2025)
23. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)

24. Woloshuk, A., Khochare, S., Almulhim, A.F., McNutt, A.T., Dean, D., Barwinska, D., Ferkowicz, M.J., Eadon, M.T., Kelly, K.J., Dunn, K.W., et al.: In situ classification of cell types in human kidney tissue using 3d nuclear staining. Cytometry Part A **99**(7), 707–721 (2021)
25. Yang, Z., Zhang, J., Luo, X., Wu, X., Lu, Z., Shen, L.: Medkan: An advanced kolmogorov-arnold network for medical image classification. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3090–3097. IEEE (2025)
26. Zhang, J., Yang, Z., Luo, X., He, S., Shen, L.: Sdc-net: Semi-supervised breast ultrasound lesion segmentation via semantic decoupling. Pattern Recognition p. 113216 (2026)
27. Zhang, J., Yang, Z., Wu, X., Luo, X., Liu, P., Shen, L.: Maximizing incremental information entropy for contrastive learning. In: The Fourteenth International Conference on Learning Representations (2026)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)